Series 10

1. In this series we explore the dataset `vehicle.dat` which can be found at

`"http://stat.ethz.ch/Teaching/Datasets/NDK/vehicle.dat"`.

The dataset contains 846 observations of 19 variables. The goal is to classify the response (which is named `Class`) into four different car types (`bus,van,saab,opel`) by means of 18 predictors such as compactness, some information about the car axes, and certain length ratios of the silhouettes of the cars. For this, we will use the CART algorithm (package `rpart`) with cost-complexity optimized size.

R-Hints: Use the R-skeleton available on the course's website.

a) Generate a classification tree using `rpart`. Use the arguments `cp = 0` and `minsplit = 30` to get a tree that is too large and overfits the data. Plot it and comment on the tree.

b) Now we prune the tree from **a)**. First, generate a cost-complexity plot and determine the size of the optimal tree according to the *one standard-error rule.* Then, use the cost-complexity table to get the `cp` value of this optimal tree and use this `cp` value to prune the tree from **a)**. Finally, visualize the pruned tree and compare it with the unprunned tree.
Note: Use the cost-complexity plot to find a range of plausible values and then use the cost-complexity table to pick the exact `cp`.
R-Hints: For this task you will need the functions `plotcp()` and `printcp()` to choose the optimal tree, and `prune.rpart()` to prune your tree.

c) Calculate the misclassification rate of the pruned tree you found in **b)**.

d) To investigate the predictive power, compute the bootstrap generalization error and the leave-one-out cross-validated performance (based on 0-1 loss) of the procedure used above. Use $B = 1000$ bootstrap-samples and set the seed for reproducibility. Comment on your results.
Note: The optimal `cp` value depends on the data that is being used to generate the tree. Therefore, you need to calculate it again every time you prune a tree, otherwise your estimations of the generalization error will be too optimistic.

e) **(optional)** Calculate the out-of-bootstrap sample generalization error (cf. Chapter 5.2.5 of the lecture notes). Compare this value with the values of the (standard) bootstrap generalization error and the cross-validation error from **d)**.

Preliminary discussion: Friday, May 20.

Deadline: Friday, May 27.